Exercise 16

Simek, Kathryn

2020-11-01

Labeled data is not always available. For these types of datasets, you can use unsupervised algorithms to extract structure. The k-means clustering algorithm and the k nearest neighbor algorithm both use the Euclidean distance between points to group data points. The difference is the k-means clustering algorithm does not use labeled data.

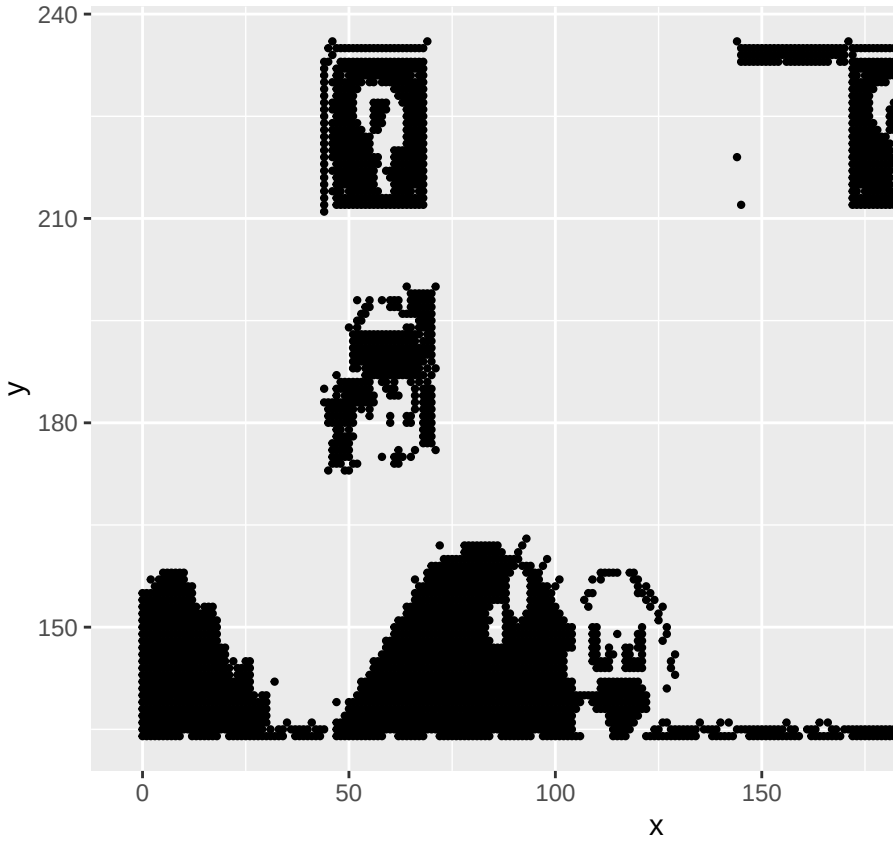
In this problem, you will use the k-means clustering algorithm to look for patterns in an unlabeled dataset. The dataset for this problem is found at data/clustering-data.csv.

```
setwd("C:/Users/katie/OneDrive/Documents/GitHub/dsc520")

mydata <- read.csv("data/clustering-data.csv")

library(ggplot2)

ggplot(mydata, aes (x=x, y=y)) + geom_point(size=.8)
```



a. Plot the dataset using a scatter plot.

b. Fit the dataset using the k-means algorithm from k=2 to k=12. Create a scatter plot of the resultant clusters for each value of k.

K=2

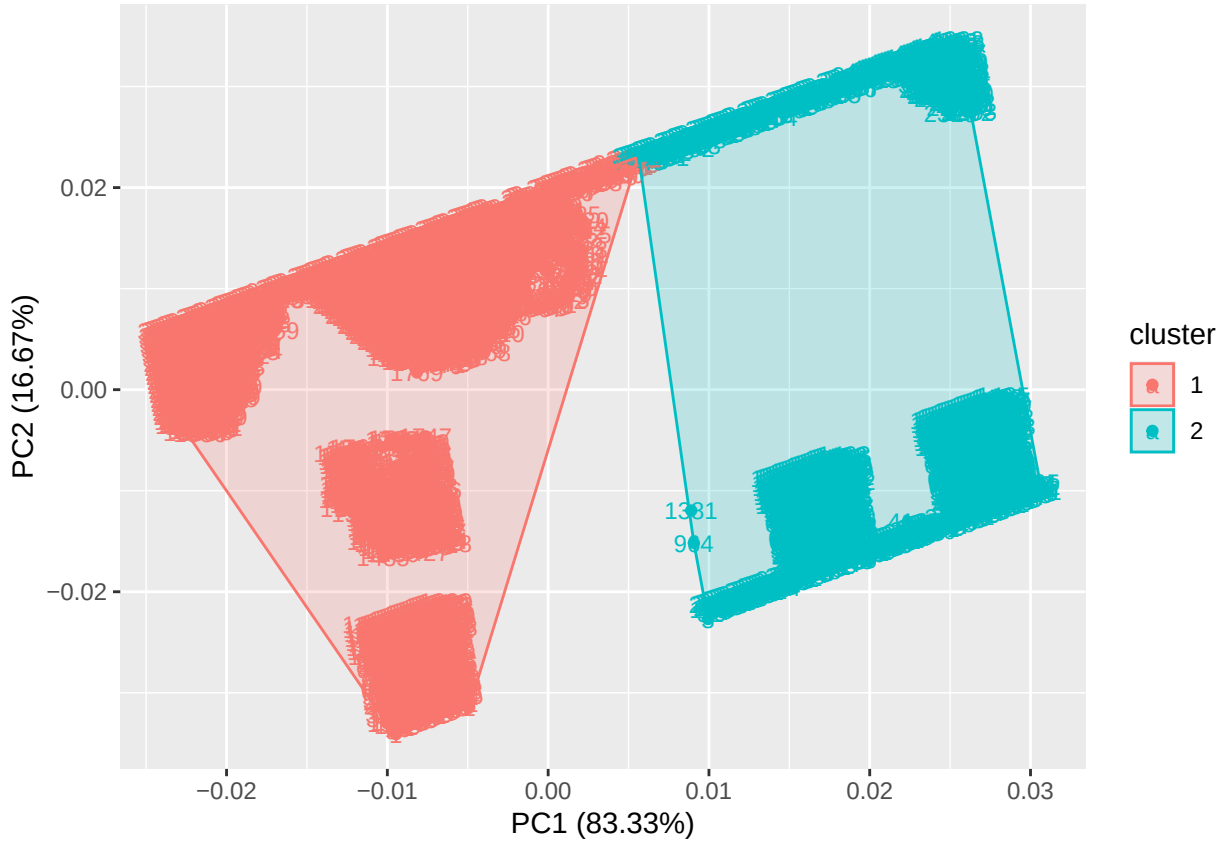
```
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 4.0.3
```

```
autoplot(kmeans(mydata, 2), data = mydata,
         label = TRUE, label.size = 3, frame = TRUE)
```

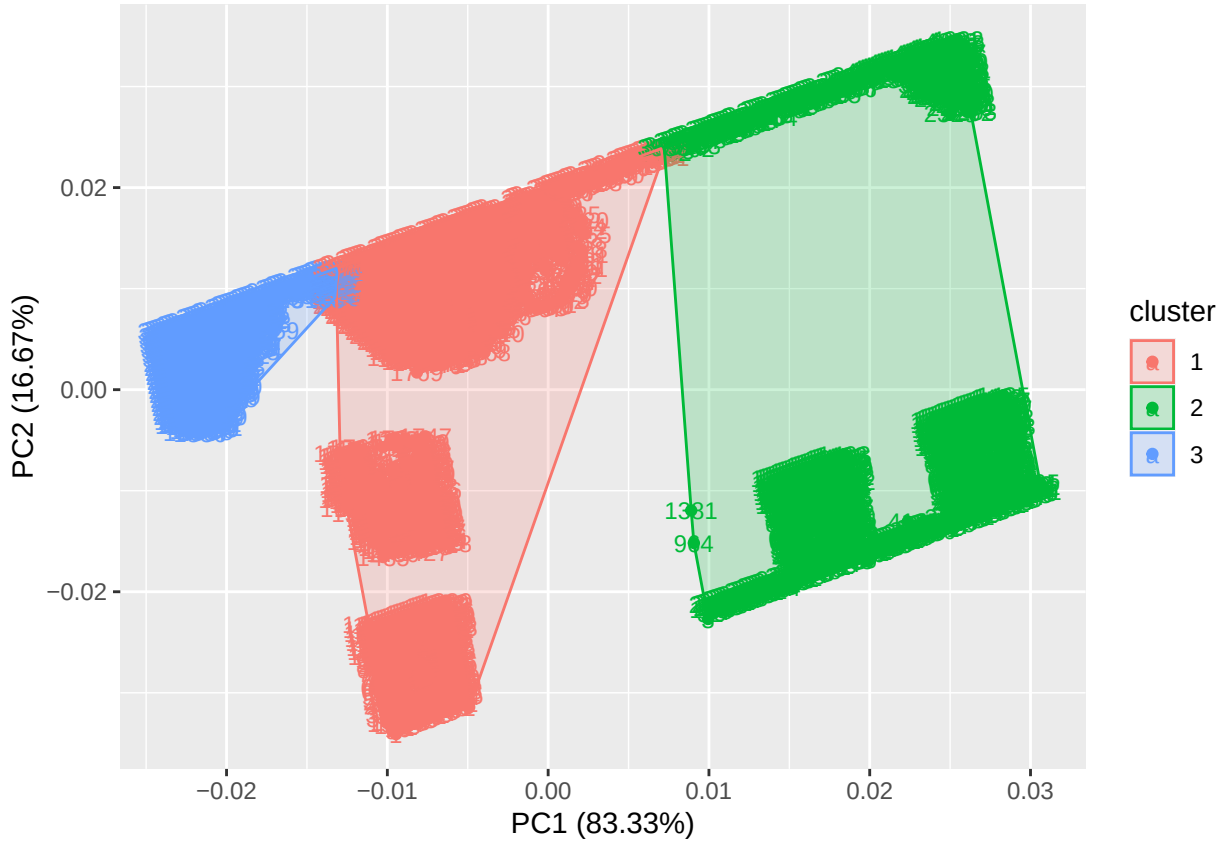
```
## Warning: `select_()` is deprecated as of dplyr 0.7.0.
## Please use `select()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.

## Warning: `group_by_()` is deprecated as of dplyr 0.7.0.
## Please use `group_by()` instead.
## See vignette('programming') for more help
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```



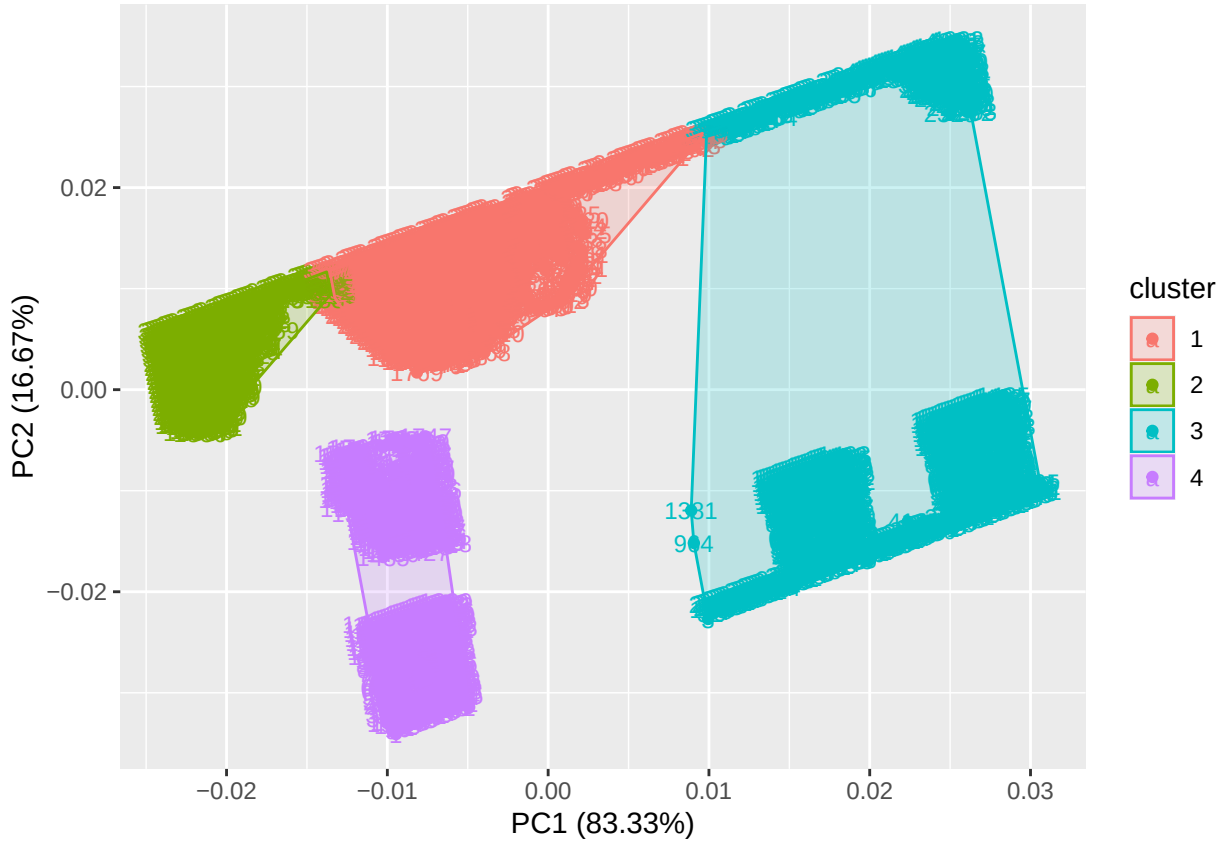
k=3

```
autoplot(kmeans(mydata, 3), data = mydata,
         label = TRUE, label.size = 3, frame = TRUE)
```



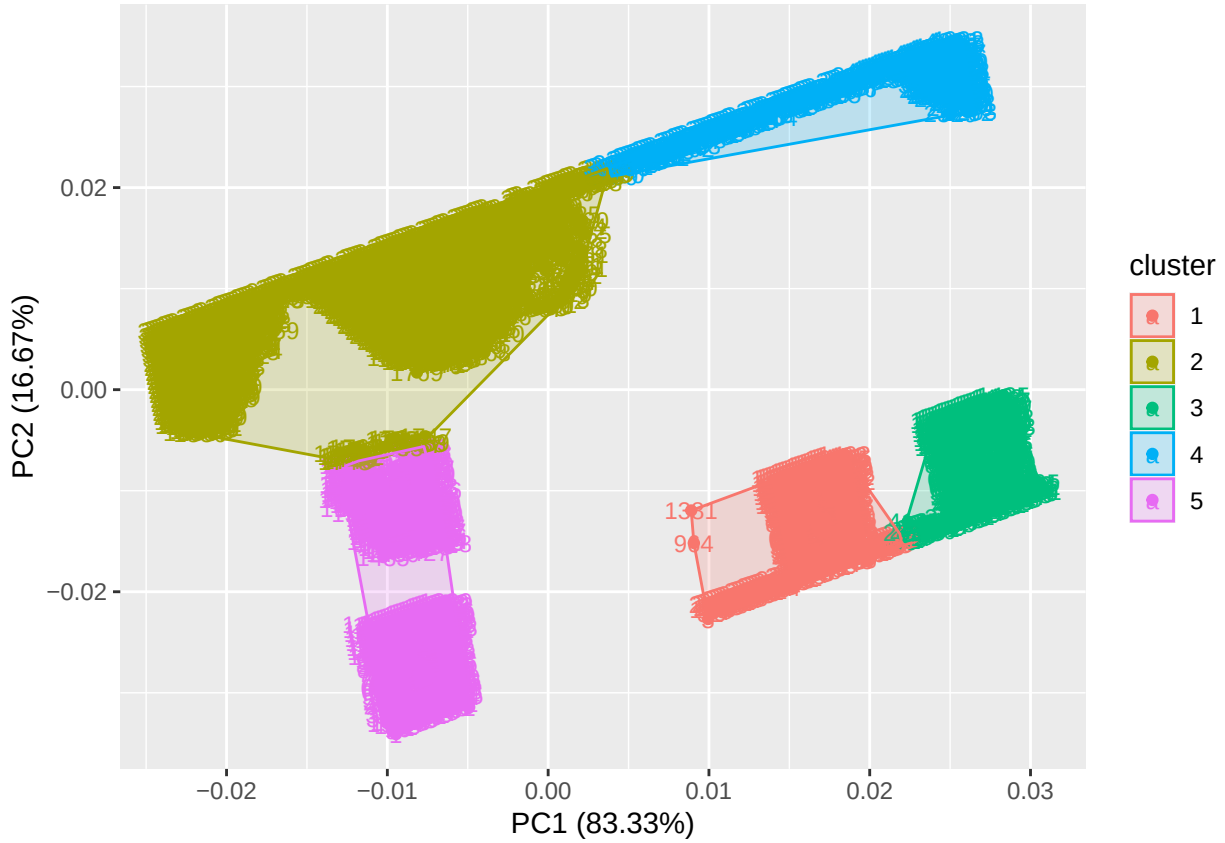
k=4

```
autoplot(kmeans(mydata, 4), data = mydata,
         label = TRUE, label.size = 3, frame = TRUE)
```



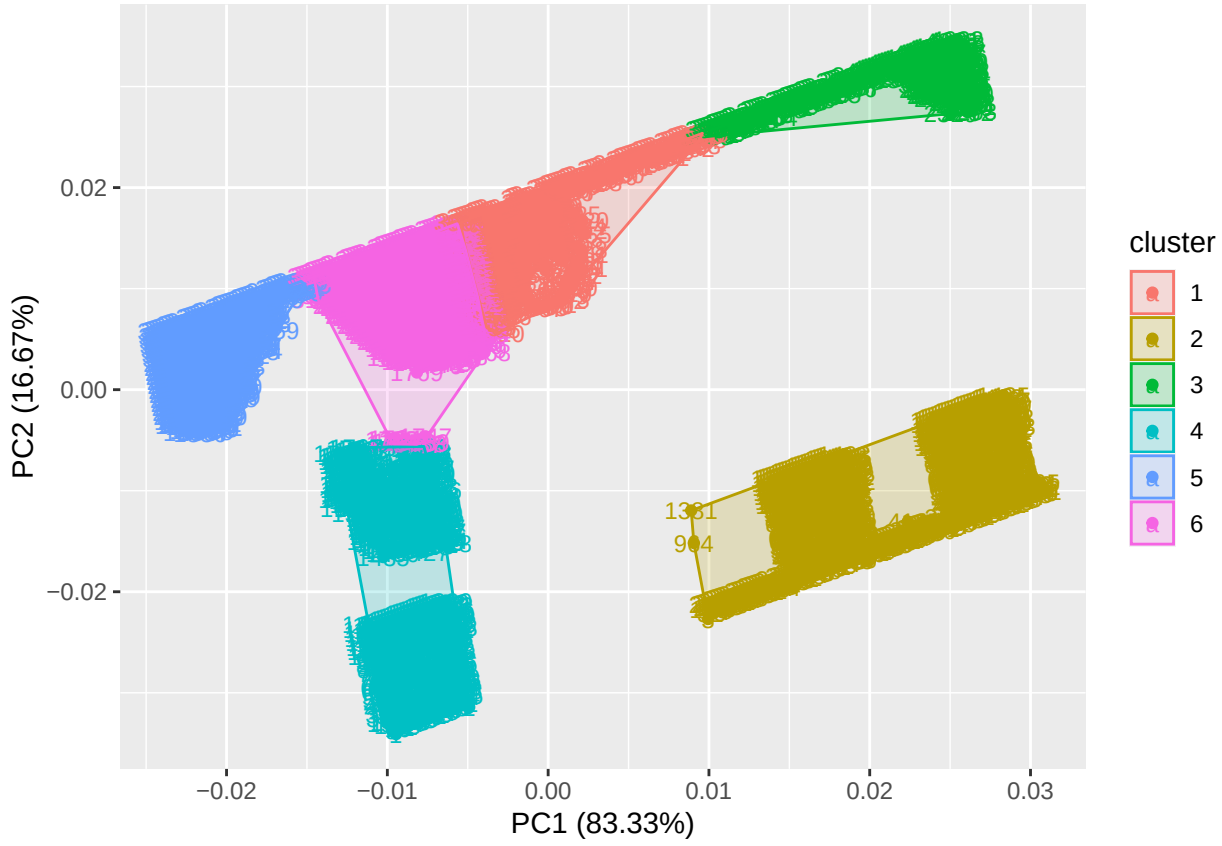
k=5

```
autoplot(kmeans(mydata, 5), data = mydata,
         label = TRUE, label.size = 3, frame = TRUE)
```



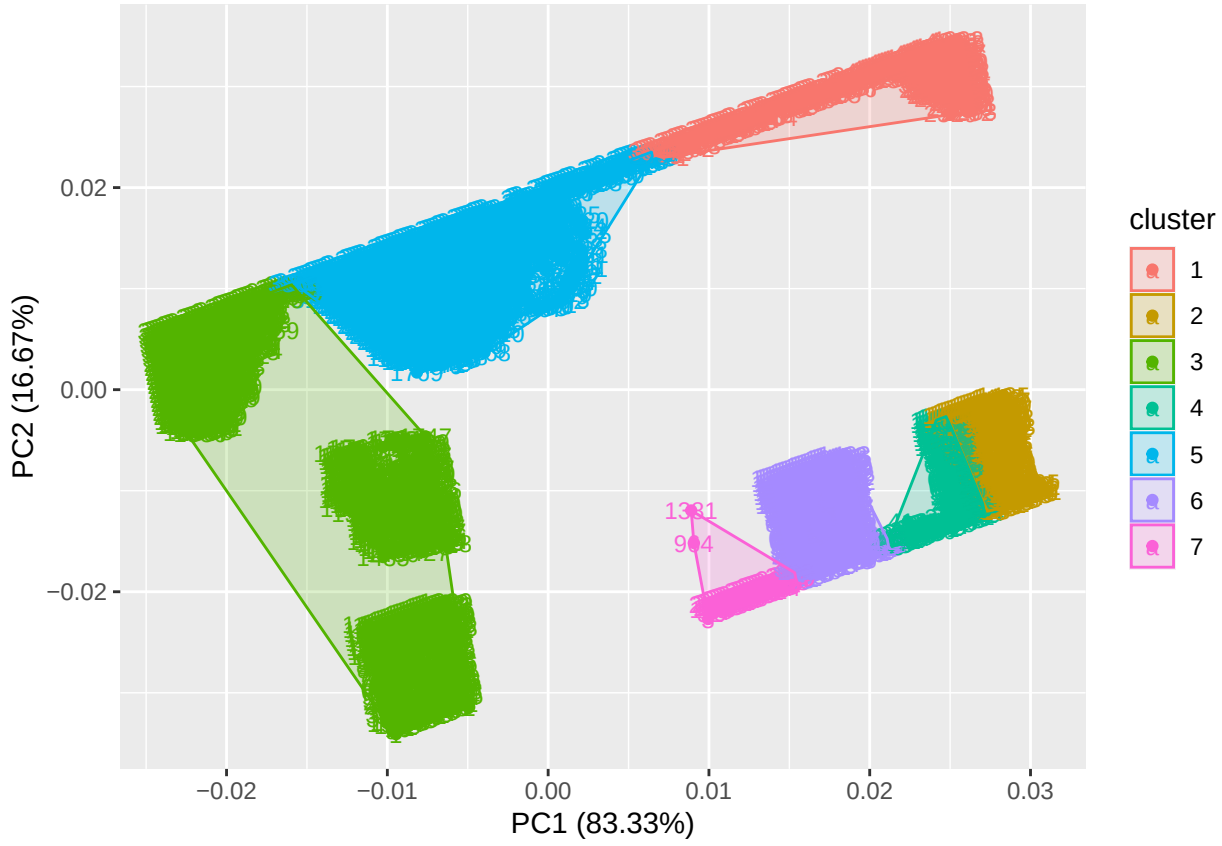
k=6

```
autoplot(kmeans(mydata, 6), data = mydata,
         label = TRUE, label.size = 3, frame = TRUE)
```



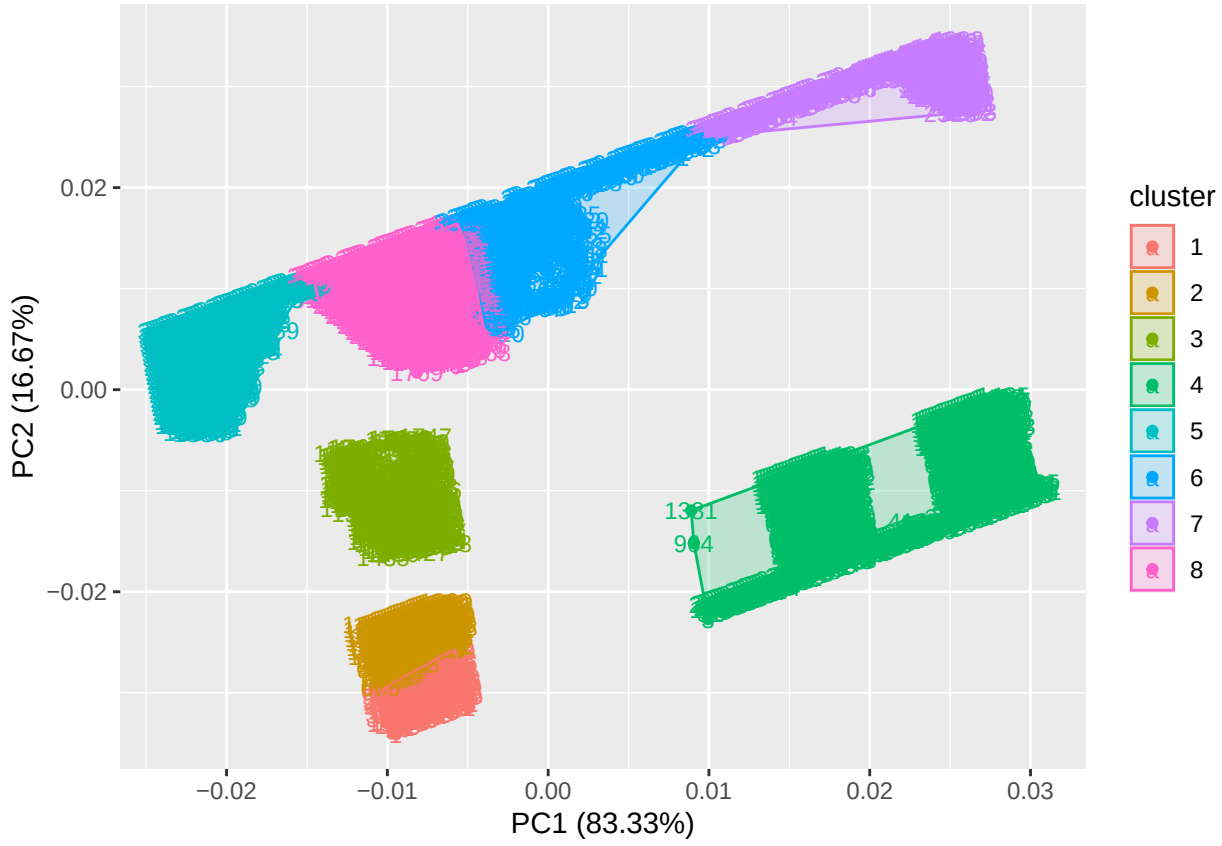
k=7

```
autoplot(kmeans(mydata, 7), data = mydata,
         label = TRUE, label.size = 3, frame = TRUE)
```



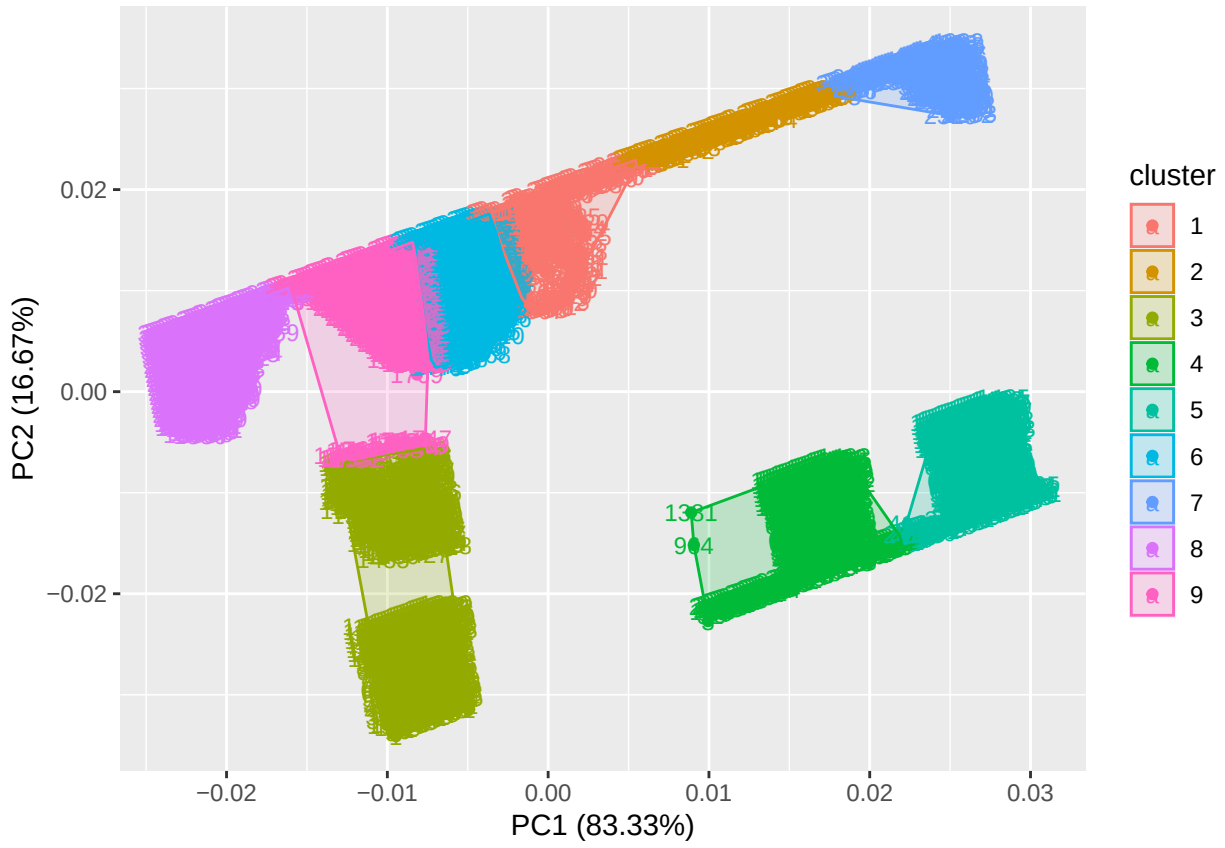
k=8

```
autoplot(kmeans(mydata, 8), data = mydata,
         label = TRUE, label.size = 3, frame = TRUE)
```



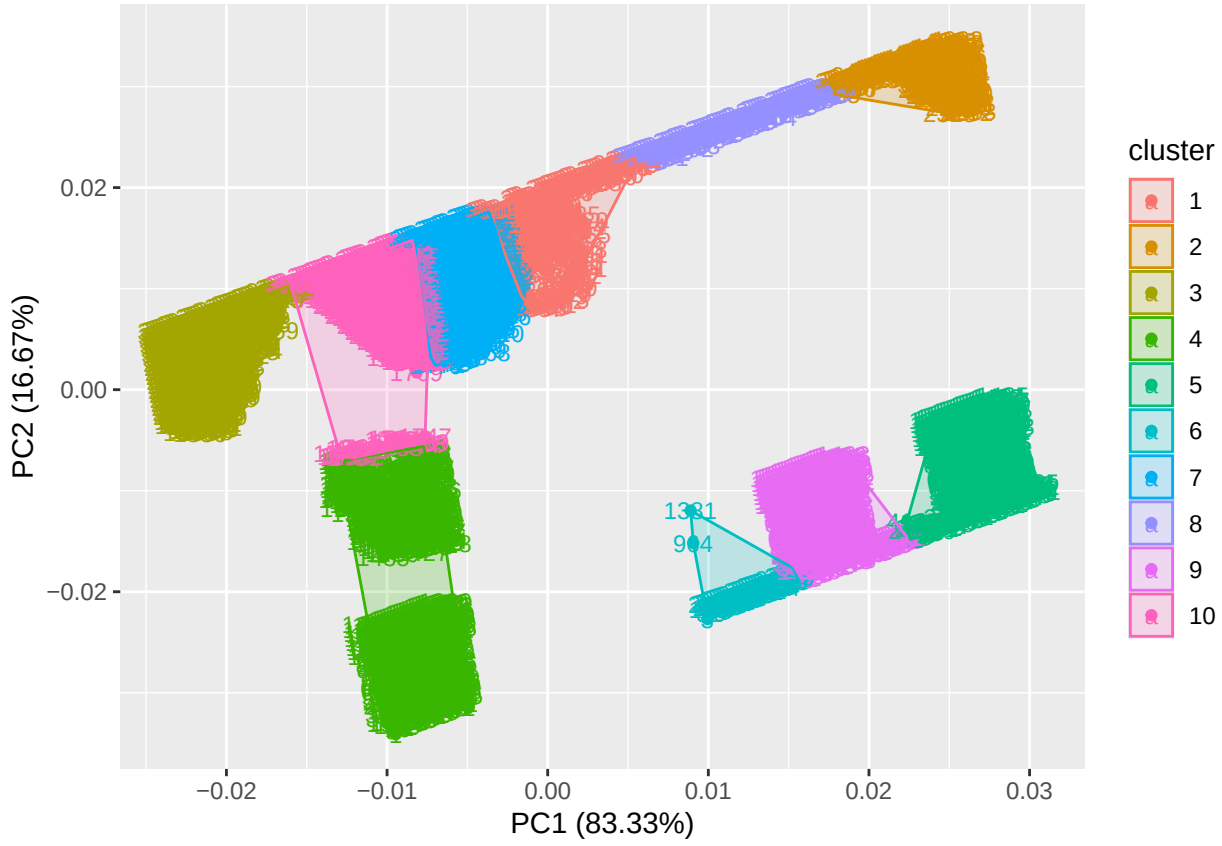
k=9

```
autoplot(kmeans(mydata, 9), data = mydata,
         label = TRUE, label.size = 3, frame = TRUE)
```



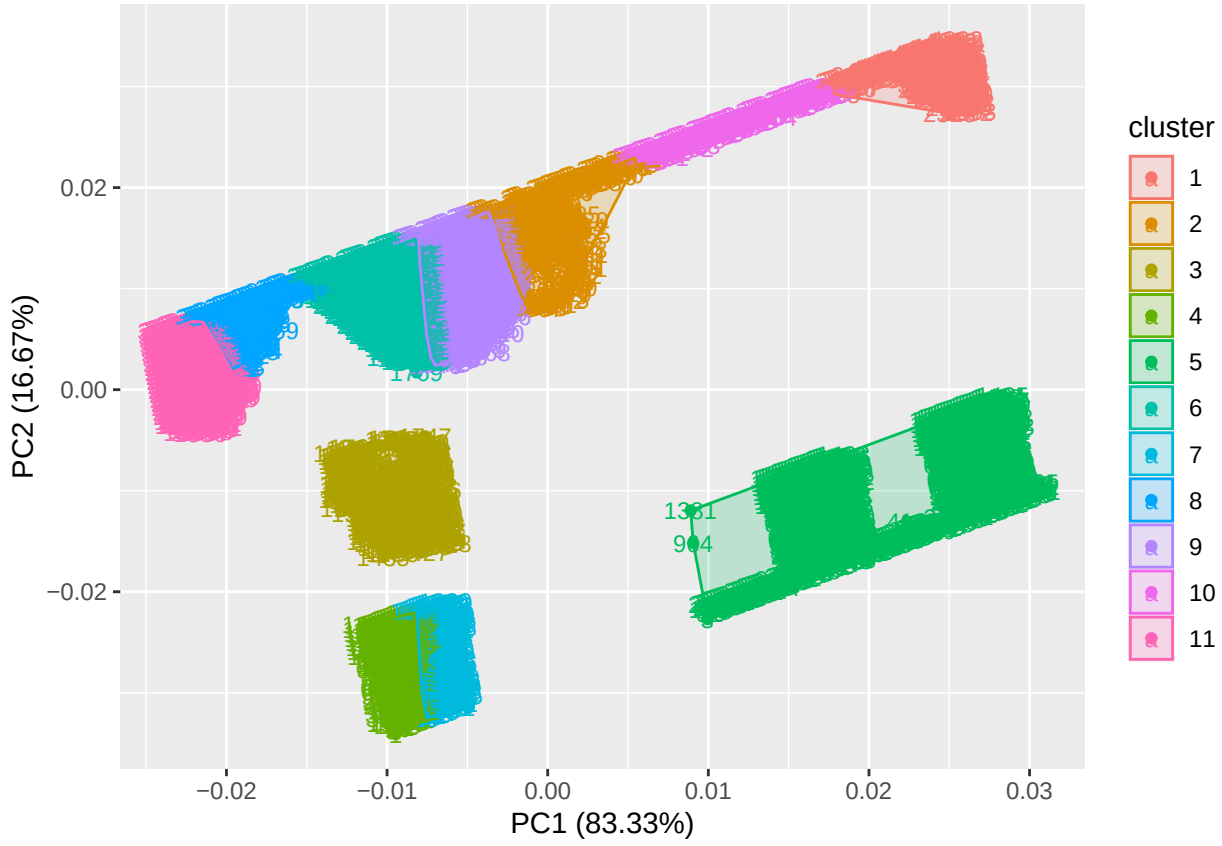
k=10

```
autoplot(kmeans(mydata, 10), data = mydata,
         label = TRUE, label.size = 3, frame = TRUE)
```



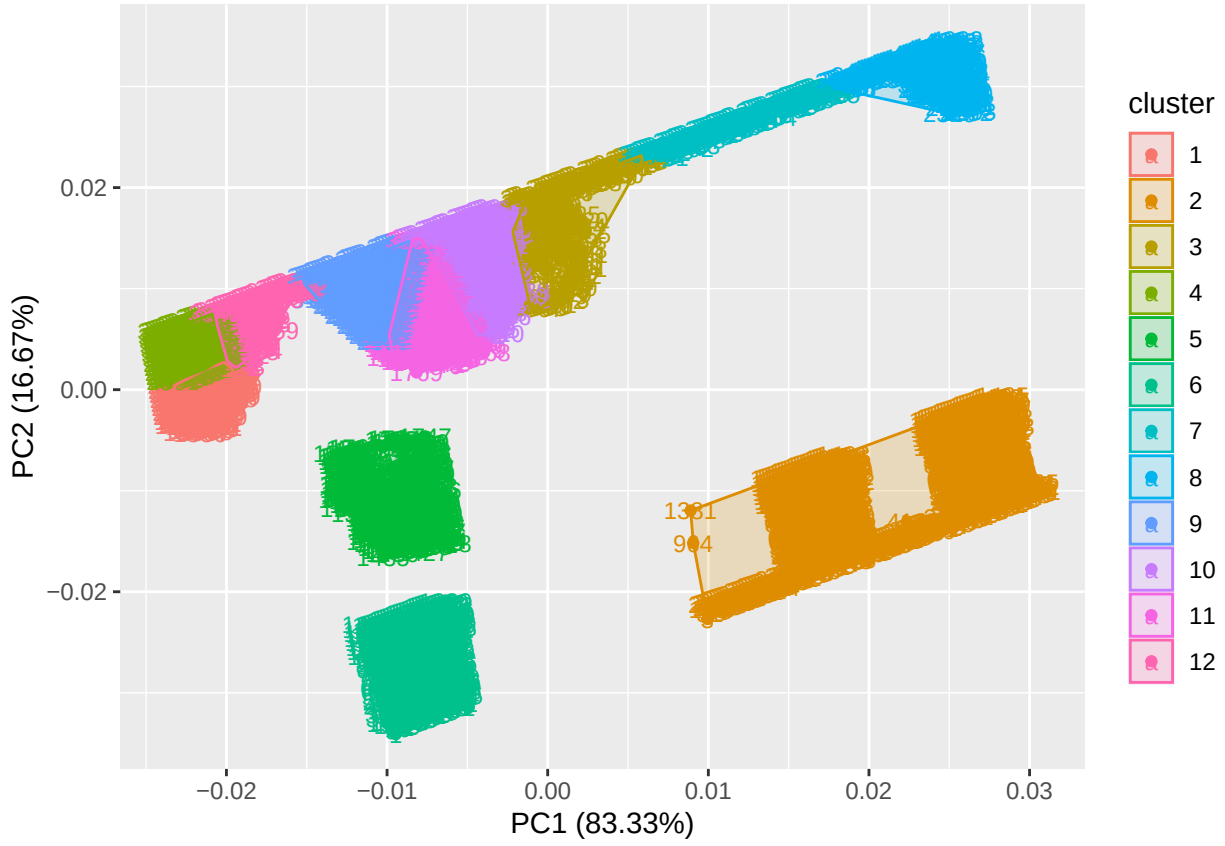
k=11

```
autoplot(kmeans(mydata, 11), data = mydata,
         label = TRUE, label.size = 3, frame = TRUE)
```



k=12

```
autoplot(kmeans(mydata, 12), data = mydata,
         label = TRUE, label.size = 3, frame = TRUE)
```

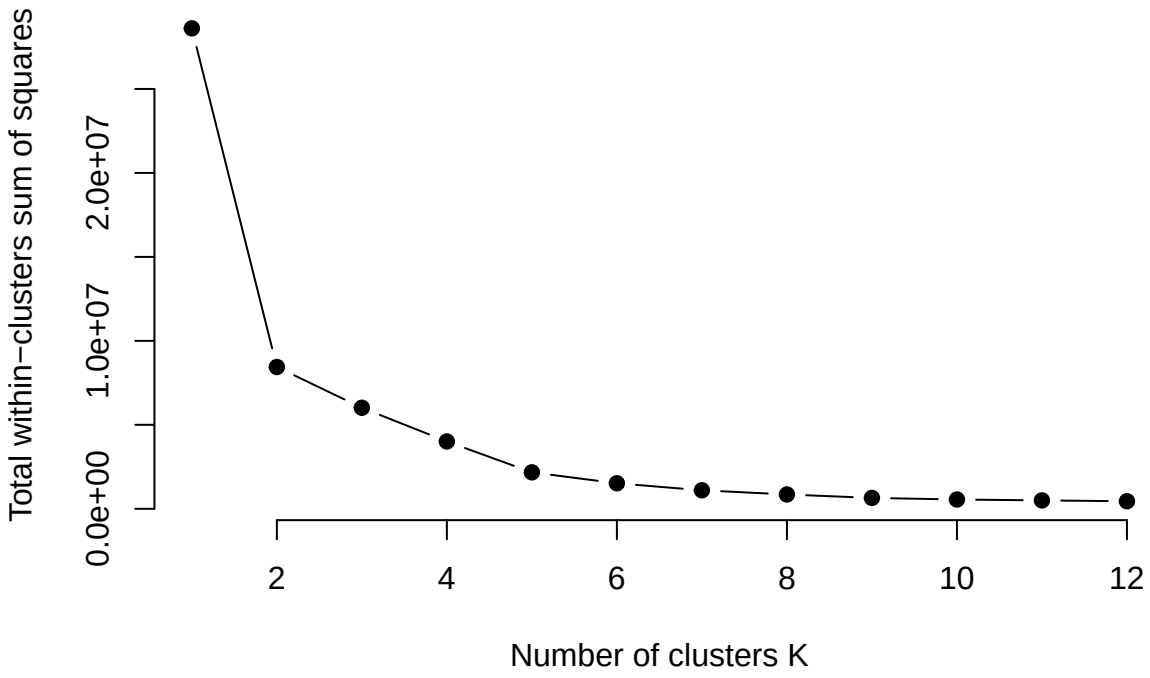


c. As k-means is an unsupervised algorithm, you cannot compute the accuracy as there are no correct values to compare the output to. Instead, you will use the average distance from the center of each cluster as a measure of how well the model fits the data. To calculate this metric, simply compute the distance of each data point to the center of the cluster it is assigned to and take the average value of all of those distances.

```
wss <- sapply(1:12,
              function(k){kmeans(mydata, k, nstart=25,iter.max = 15)$tot.withinss})

plot(1:12, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```

Calculate this average distance from the center of each cluster for each value of k and plot it as a line chart where k is the x-axis and the average distance is the y-axis.



One way of determining the "right" number of clusters is to look at the graph of k versus average distance and finding the "elbow point". Looking at the graph you generated in the previous example, what is the elbow point for this dataset? Answer: The elbow point is at 5 clusters.